3 Panorama de los corpus y textos del portugués europeo contemporáneo

Abstract: El presente capítulo tiene como primer objetivo proporcionar una herramienta que permita al estudioso (i) determinar fácil y rápidamente los corpus que le son útiles para realizar una investigación de corpus sobre cualquier tema relacionado con el portugués europeo (PE) contemporáneo, (ii) saber cómo consultarlos, (iii) encontrar estudios que describan más detalladamente las aplicaciones de estos corpus (sección 1). La selección de los corpus disponibles se ha hecho según algunos criterios predefinidos. Incluimos los proyectos concluidos o en gran medida ya disponibles, que contienen principalmente muestras de la lengua contemporánea (segunda mitad del siglo XX-siglo XXI). Además, los textos son de libre y fácil acceso (en forma digital) y su tamaño es razonablemente grande.

El segundo objetivo de este capítulo es ilustrar la variación tipológica de los géneros textuales y de las grandes zonas dialectales del PE actual a través de unas muestras representativas (sección 2). Pese a ser imposible dar un elenco exhaustivo de distintos géneros y variedades dentro de los márgenes de este artículo, tratamos así de dar una idea global de la variación que presenta el PE hoy en día.

Keywords: corpus, portugués europeo contemporáneo, géneros textuales, dialectos

1 Corpus

Distinguimos dos grandes grupos de corpus: los corpus disponibles para consulta en la Red (1.1) y los corpus de textos orales disponibles en su totalidad, pero sin acceso «directo» vía un motor de búsqueda online (1.2). Subdividimos el primer grupo en los corpus de medios mixtos (1.1.1), escritos (1.1.2), orales (1.1.3) y, por fin, los corpus manualmente anotados con objetivo específico (1.1.4).

Cada corpus se presenta sistemáticamente según el esquema siguiente: ofrecemos un listado con información puntual sobre (i) el acceso y el soporte del corpus (ii) el tamaño del corpus (iii) el/los autores (iv) la variante geográfica. A pesar de tratarse de corpus del PE, incluimos el punto (iv) para dar cuenta de los corpus que contienen también textos de otras variantes geográficas. Después, se describe con algo más de detalle el tipo de textos incluidos, la periodización, la anotación y las modalidades de búsqueda.

1.1 Corpus disponibles para consulta en la Red

Los corpus descritos en este apartado son consultables mediante un buscador online que permite crear órdenes de búsqueda de distintos tipos. Sin embargo, los corpus marcados con * se pueden obtener también en su totalidad (por ej. por descarga).

1.1.1 Medios Mixtos

Corpus de Referência do Português Contemporâneo, parte escrita (CRPC)*

- Partes escritas del CRPC accesibles en <http://alfclul.clul.ul.pt/CQPweb/>; acceso con registro en <http://www.clul.ul.pt/pt/recursos/183-crpc#cqp>
- 290 millones de palabras para la parte del PE escrito; 1,4 millones para la parte oral
- Centro de Lingüística de la Universidad de Lisboa (CLUL)[1]
- Mayoritariamente PE; contiene también otras variantes nacionales

El CRPC constituye el mayor corpus representativo del PE, de distintos géneros, que sobre todo para el lenguaje escrito ofrece una colección enorme de textos (Bacelar do Nascimento 2000; Généreux/Hendrickx/Mendes 2012). Reúne varios géneros escritos (literario, prensa, técnico, científico, didáctico, folletos, decisiones del Tribunal Supremo de Justicia, sesiones parlamentarias), de la segunda mitad del siglo XIX hasta el año 2006, pero con una mayoría de textos posteriores a 1970. El corpus escrito está lematizado y tiene etiquetaje morfosintáctico POS. El buscador permite el uso de comodines y el estudio de colocaciones. Así se pueden indagar partes de palabras, secuencias de palabras, lemas, y categorías gramaticales. El buscador presenta además una función para bajar los resultados en un fichero (por ej. excel), y, para usuarios registrados, guardarlos y categorizarlos. No es posible hacer búsquedas según la cronología (por ej. solo textos posteriores al año 2000).

La parte oral del corpus consiste en una serie de subcorpus que trataremos más adelante: *Português Fundamental*, *Português Falado*, *C-ORAL-ROM*.

Dos subcorpus escritos del CRPC están disponibles de forma completa con anotación en el catálogo de ELRA (European Language Resources Association):

a) Corpus Literário (Corpus LT)

- <http://catalog.elra.info/product_info.php?products_id=1178> (gratuito para investigadores)
- 1,7 millones de palabras
- PE / PB

El *Corpus LT* comprende 70 textos de autores renombrados de lengua ptg. de los años 1810 hasta 1940. Se presenta en forma lematizada y anotada con etiquetas morfosintácticas POS (anotación automática).

b) Corpus de Política (PTParl)

- <http://catalog.elra.info/product_info.php?products_id=1179> (gratuito para investigadores)
- 1 millón de palabras
- PE

[1] Como se verá adelante, el CLUL pone a disposición varios corpus en su sitio web: <http://www.clul.ul.pt/en/resources>. Además de las fuentes que se citan para cada corpus en cuestión, aconsejamos también consultar esta plataforma para más información técnica sobre los corpus.

El *PTParl* fue automáticamente lematizado y anotado con etiquetas morfosintácticas POS. Contiene las actas del parlamento ptg. tal como aparecen en el *Diário da Assembleia da República*.

Corpus do Português

- <http://www.corpusdoportugues.org/>
- 10 millones de palabras para la parte PE del siglo XX
- Mark Davies (Brigham Young University) y Michael J. Ferreira (Georgetown University)
- PE y portugués brasileño (PB)

El *Corpus do Português* es un corpus de textos escritos y orales que comprende los siglos XIII-XX, construido a base de textos escaneados, textos accesibles en internet y a base de otros corpus (CORDIAL-SIN, Corpus de Linguateca, etc.) y bases de datos textuales y de textos escaneados (Davies 2014). Para el siglo XX, el corpus contiene textos de ficción, periodísticos, académicos y orales de distintos tipos no especificados. Los textos escritos del siglo XX abarcan alrededor de 3 millones de palabras por género, y los textos orales cerca de 1 millón de palabras. Se trata de un corpus lematizado, con etiquetaje POS, que permite el uso de comodines y la búsqueda de colocaciones. La interfaz de búsqueda deja visualizar los resultados por siglo, y por registro para el siglo XX. Se pueden crear y guardar listas personalizadas. No se posibilitan búsquedas de períodos más precisos (dentro de un mismo siglo) o por autor, y tampoco búsquedas por variedad geográfica y por género al mismo tiempo. Para los textos orales no hay acceso a las grabaciones. El buscador tiene la misma arquitectura básica que el Corpus del Español (### 7 Panorama de los corpus y textos del español peninsular contemporáneo).

Análise Contrastiva de Variedades do Português (VARPORT)

- <http://www.letras.ufrj.br/varport/>
- Número de palabras desconocido
- Universidade Federal do Rio de Janeiro (UFRJ) y CLUL
- PE / PB

Corpus de textos escritos (anuncios, editoriales, noticias) y orales del siglo XX cuya parte del PE se basa en algunos corpus desarrollados por el CLUL, a saber el CRPC (véase *supra*) y el Portugués Fundamental (*infra*) para la parte oral. El VARPORT ambiciona ofrecer un corpus de dimensión, distribución y arquitectura comparables para las variantes PE y PB. Para la parte oral hay acceso a la grabación audio, y a la metainformación (registro, edad, sexo, etc.). Sin embargo, el formato de los ficheros en línea es difícil de procesar por la falta de un motor de búsqueda. El corpus es de dimensión reducida.

1.1.2 Escrito

CETEMPúblico*

- <http://www.linguateca.pt/CETEMPublico/>; también accesible para descarga de forma completa a petición
- 190 millones de palabras
- Linguateca[2]
- PE

Este corpus escrito incluye 2.600 ediciones del periódico *Público* (1991-1998). El corpus está accesible para búsquedas online a través del proyecto AC/DC (*Acesso a Corpos / Disponibilização de corpos*), en la plataforma de *Linguateca* y a petición está disponible de forma íntegra para descarga (Rocha/Santos 2000). Los artículos están subdivididos en extractos de, como máximo, unas frases. Por sus largas dimensiones y por las dos maneras de acceso, el CETEMPúblico ha sido utilizado frecuentemente por lingüistas interesados en estudiar fenómenos lingüísticos del portugués escrito contemporáneo, pero también por estudiosos del procesamiento del lenguaje natural en el marco del desarrollo de herramientas. El corpus presenta varios niveles de anotación (categoría morfosintáctica POS, flexión verbal y nominal, lematización y etiquetaje sintáctico (mediante el programa PALAVRAS)). Esta anotación ayuda a hacer búsquedas de (partes de) palabras, secuencias de palabras, etiquetas POS, flexión verbal y nominal, lema, constituyentes sintácticos, funciones sintácticas y la distribución de estas informaciones en el corpus.

Avante

- <http://www.linguateca.pt/acesso/corpus.php?corpus=AVANTE>
- 6,8 millones de palabras
- Linguateca
- PE

El corpus Avante contiene textos de los años 1997-2002 de la revista semanal *Avante!*, órgano oficial del *Partido Comunista Portugués*. Tal como el CETEMPúblico (*supra*), se pone a disposición a través del proyecto AC/DC, con los mismos niveles de etiquetaje y posibilidades de búsqueda idénticas. Además, se pueden hacer búsquedas en función de la edición de la revista.

Natura/Minho

- <http://www.linguateca.pt/acesso/corpus.php?corpus=NATMINHO>
- 1,7 millones de palabras
- Linguateca y Universidade do Minho
- PE, región del Miño

---

[2] La plataforma digital *Linguateca* ofrece una serie de corpus y recursos para el procesamiento computacional del idioma ptg. A continuación se mencionan varios de estos corpus, accesibles a través del proyecto AC/DC <http://www.linguateca.pt/ACDC/>.

El corpus *Natura/Minho* recopila textos del periódico regional ptg. *Diário do Minho*. Está disponible también en la plataforma del proyecto AC/DC y presenta las mismas modalidades de búsqueda que el CETEMPúblico (cf. *supra*).

CONDIVport

- <http://www.linguateca.pt/acesso/corpus.php?corpus=CONDIV>
- 5,6 millones de palabras, de los cuales 3,3 millones para el PE
- Linguateca y Universidade do Minho
- PE / PB

El *CONDIVport* se compone de textos de las décadas 1950, 1970 y 2000, extraídos de periódicos y revistas ptg. y bras. Los textos abarcan tres campos semánticos: fútbol, vestuario y moda, salud. La parte que está accesible a través del proyecto AC/DC de Linguateca, forma parte de un corpus más extenso CONDIVport (Silva 2008). Este corpus se constituyó con el objetivo de investigar si, desde los años 1950, las dos variantes nacionales del ptg. se caracterizan por un proceso de convergencia o divergencia léxica. El acceso en internet ofrece las mismas opciones de búsqueda que las del CETEMPúblico (cf. *supra*). Además se posibilita hacer búsquedas por variante del ptg. y por campo semántico.

CoNE

- <http://www.linguateca.pt/acesso/corpus.php?corpus=CONE>
- 675.000 palabras
- Linguateca
- PE / PB

El corpus *CoNE* (*Correio Não Endereçado*) consta de correos electrónicos publicitarios o informativos, recibidos por miembros del equipo de Linguateca entre 2001 y 2006. Está disponible a través de la plataforma AC/DC, con los mismos niveles de etiquetaje y posibilidades de búsqueda que el CETEMPúblico.

Corpus PostScriptum - FLY*

- <http://alfclul.clul.ul.pt/cards-fly/index.php?page=mainen>
- 2000 cartas
- CLUL
- Cartas en ptg. de distintos orígenes geográficos

El corpus *PostScriptum – FLY* constituye una colección de cartas escritas en ptg., producidas entre 1900-1975 en la esfera privada de autores de varios estratos sociales, en un contexto de guerra, emigración,

prisión o exilio. Los textos tienen el doble formato de corpus lingüístico y de edición crítica puesta a disposición en internet y comentada desde una perspectiva histórica, lingüística y sociológica. Las cartas están anonimizadas, se presentan en formato XML-TEI y están anotadas con etiquetas textuales y palabras-clave sociológicas. Cada una de las cartas puede descargarse en formato XML o PDF. La totalidad del corpus puede indagarse en función del año, del sitio de emisión, del tipo de carta (amor, amistad, noticias, etc.), de palabras-clave predefinidas o mediante una búsqueda libre de palabras.

Corpus paralelo bidireccional de português e inglês (COMPARA)

- <http://193.136.2.104/COMPARA/index.php>
- 1,1 millones de palabras para el PE original y traducido (en total 1,4 millones de palabras para el ptg.)
- Linguateca
- Sobre todo PE; contiene también algunas otras variantes nacionales

El *COMPARA* es un corpus paralelo de textos literarios traducidos del y al inglés de los siglos XIX-XX, alineados por frase. El corpus está lematizado y etiquetado morfosintácticamente (POS). El buscador permite el uso de comodines, y posibilita limitar la búsqueda a secciones bien precisas del corpus (por ej. limitación cronológica, por autor, por variante, por lengua original/traducida, etc.). Se pueden buscar (secuencias de) (partes de) palabras según varios parámetros. Se obtiene un contexto de más o menos una frase.

1.1.3 Oral

Projecto Corpus Dialectal para o Estudo da Sintaxe (CORDIAL-SIN)*

- <http://www.clul.ul.pt/pt/recursos/226-corpus-syntax-oriented-corpus-of-portuguese-dialects-cordial-sin>
- 600.000 palabras
- CLUL
- PE

El *CORDIAL-SIN* es un corpus oral que contiene grabaciones de discurso libre y semi-dirigido hechas en varias localidades de Portugal (cf. Carrilho 2010). Están a disposición cuatro formatos con metainformación sobre los informantes: 1) una transcripción conservadora (solamente disponible en formato PDF) con información sobre marcas de la oralidad como pausas, superposiciones en la producción, hesitaciones, reformulaciones, repeticiones, formas truncadas, variantes fonéticas, etc.; 2) una transcripción ortográfica normalizada sin marcas de oralidad; 3) la transcripción ortográfica con anotación morfosintáctica (etiquetaje POS e información flexiva); 4) la transcripción ortográfica con anotación sintáctica (por oración), actualmente solo para textos de 14 localidades (este fichero puede indagarse con la herramienta *CorpusSearch*). Los varios formatos se ajustan a distintos tipos de búsqueda, con manuales

claros sobre la transcripción y anotación. Los textos del CORDIAL-SIN forman una muestra representativa de las grabaciones (disponibles en el CLUL a petición) reunidas en el marco de varios proyectos de atlas lingüísticos: el *Atlas Linguístico e Etnográfico de Portugal e da Galiza (ALEPG)*, el *Atlas Linguístico do Litoral Português*, el *Atlas Linguístico e Etnográfico dos Açores*, y el proyecto *Fronteira Dialectal do Barlavento Algarvio*[3].

Corp-Oral*

- <http://www.iltec.pt/spock/>; accesible a petición para descarga <http://corpus1.mpi.nl/ds/imdi_browser/>
- 56 grabaciones (50 h)
- Instituto de Linguística Teórica e Computacional (ILTEC)
- PE

El Corp-Oral es un corpus oral de diálogo libre de hablantes entre 12 y 74 años del área metropolitana de Lisboa (ILTEC/FCT 2012). De las 50 horas de grabación, 30 horas están transcritas ortográficamente. Sociolingüísticamente, los hablantes representan diversos niveles académicos y profesionales, así como relaciones más o menos cercanas entre sí. En consecuencia, se presentan diálogos de distintos grados de formalidad. Las normas de transcripción ortográfica siguen en gran medida las normas del C-ORAL-ROM, con indicaciones paralingüísticas (repeticiones, interrupciones, pausas, etc.). La aplicación online (*Spock*) para la visualización de los datos permite ver ocurrencias de palabras y secuencias de palabras en contexto, y escuchar la grabación correspondiente. Sin embargo, para mejor calidad de las grabaciones se aconseja la descarga del corpus completo a partir del *Isle MetaData Initiative* (Max Planck Institute for Psycholinguistics), tras petición de contraseña a fabiola.santos@iltec.pt.

Corpus Museu da Pessoa

- <http://www.linguateca.pt/acesso/corpus.php?corpus=MUSEUDAPESSOA>
- 1,4 millones de palabras
- Museu da Pessoa/ Núcleo Português do Museu da Pessoa
- PE / PB

El *Corpus Museu da Pessoa* reúne entrevistas transcritas (107 para el PE, 106 para el PB) posteriores al año 2000. Presenta lematización y permite el uso de comodines. Las transcripciones incluyen etiquetaje gramatical (POS) y anotación sintáctica. Está disponible en la misma plataforma de herramientas del corpus CETEMPúblico (cf. *supra*). No hay acceso a las grabaciones audio y se obtiene apenas una frase de contexto.

---

[3] Cf. http://www.clul.ul.pt/en/research-teams/516-related-projects

Rede de Difusão Internacional do Português: rádio, televisão e imprensa (ReDIP)

- <http://www.iltec.pt/?action=concord>
- 330.000 palabras
- Instituto de Linguística Teórica e Computacional (ILTEC), en colaboración con el CLUL y la Universidade Aberta
- PE

El *ReDIP* es un corpus oral y escrito compuesto de 36 textos radiofónicos y televisivos y de textos de prensa, que se dividen por temas: actualidad, ciencia, cultura, economía, deporte y opinión. El buscador en línea permite consultar solo la parte oral del corpus, recuperando (partes de) palabras en un contexto de unas líneas.

1.1.4 Corpus manualmente anotados con objetivo específico

CINTIL – Corpus Internacional do Português*

- <http://cintil.ul.pt>; también en venta en el catálogo ELRA: <http://catalog.elra.info/product_info.php?products_id=1102>
- 1 millón de palabras
- Grupo Natural Language and Speech (NLX, Universidad de Lisboa) y CLUL
- PE

El CINTIL constituye un corpus de textos escritos (34% noticias, 17% ficción, 7% otros) y orales (42%, distintos registros y situaciones comunicativas). Los textos de ficción datan de los años 1844-1997; los demás textos son posteriores a los años 1970 hasta los años 2000 inclusive. El corpus está lematizado y anotado con etiquetas morfosintácticas (POS), que contienen información sobre la flexión verbal y nominal (cf. Barreto et al. 2006). Lleva además un etiquetaje específico de dos tipos: 1) locuciones adverbiales y locuciones que pertenecen a categorías gramaticales cerradas (conj., dem., pron., interjecciones, etc.) 2) entidades nombradas (personas, localizaciones, organizaciones, obras y otros). El buscador permite un acceso fácil a los datos, y posibilita búsquedas avanzadas de (secuencias de) palabras y/o partes de palabras mediante el uso de comodines y el etiquetaje, que pueden restringirse según el género textual. Sin embargo, no se pueden delimitar periodos más específicos y tampoco acceso a la referencia exacta de la fuente de cada ejemplo. Aun así, a través de compra en ELRA se puede obtener el corpus completo para búsquedas libres.

Corpus PAROLE

- <http://www.elda.org/catalogue/en/text/W0024.html>
- 250.000 palabras
- CLUL

– PE

El PAROLE, corpus escrito y anotado (POS), está compuesto de textos periodísticos y literarios extraídos del CRPC (cf. *supra*). Fue creado en el ámbito de un proyecto europeo (LE-PAROLE[4]), que incluye cerca de 20 lenguas europeas. Para cada una de estas lenguas fue compilado un corpus de 20 millones de palabras, de las cuales unas 250.000 fueron etiquetadas con información morfosintáctica (etiquetaje automático junto con desambiguación manual), según un sistema de etiquetaje uniforme para todas las lenguas, que incluye la categoría gramatical general (N, V, etc.) y la flexión verbal y nominal (Bacelar do Nascimento et al. 1998). Como tal, este proyecto se presta a hacer estudios multilingües con base en estos corpus anotados de manera uniforme. El PAROLE sirvió de corpus de entrenamiento para el desarrollo de etiquetadores morfosintácticos para el ptg. y dio origen al corpus CINTIL (cf. *supra*).

Corpus CD Harem*

- <http://www.linguateca.pt/acesso/corpus.php?corpus=CDHAREM>; disponible para descarga en <http://www.linguateca.pt/HAREM/>
- 225.000 palabras
- Linguateca
- PE / PB

El *CD Harem* forma un conjunto de textos con etiquetaje de Entidades Mencionadas (i.e., nombres propios de personas, sitios etc.), extraídos de las colecciones doradas usadas en el proyecto HAREM (proyecto de evaluación de sistemas de Reconocimiento de Entidades Mencionadas) (Rocha/Santos 2007). El corpus entero está disponible para descarga en formato XML pero puede también indagarse online, a través del proyecto AC/DC: además de las posibilidades de búsqueda comunes con el CETEMPúblico, se permite la búsqueda en función de la Entidad Mencionada, y por variedad geográfica del ptg. (PE o PB).

Corpus Floresta*

- <http://www.linguateca.pt/acesso/corpus.php?corpus=FLORESTA>; descarga en <http://www.linguateca.pt/Floresta/levantamento.html>
- 6,7 millones de palabras
- Linguateca y Visual Interactive Syntax Learning (Universidad de Dinamarca del Sur)
- PE / PB

El corpus *Floresta* contiene 261 mil frases sintácticamente analizadas, con etiquetas morfosintácticas (POS), y de flexión y lema (Freitas/Afonso 2008). Está dividido en 4 subconjuntos (*Bosque*, *Floresta virgem*, *Selva* y *Amazônia*), cada uno con una constitución interna distinta y diferentes niveles de revisión

[4] Comisión Europea – DGXIII, Telematics Application of Common Interest – Contracto LE2 – 4017.

de etiquetaje. El conjunto *Bosque* fue completamente revisado por lingüistas y consiste de 9368 frases, provenientes de los primeros extractos de los corpus CETENFolha (PB) y CETEMPúblico (PE). La Floresta Sintáctica puede ser consultada a través del proyecto AC/DC o mediante la interfaz de búsqueda en árboles sintácticos *Milhafre* (<http://www.linguateca.pt/Floresta/milhafre/>), pero puede también descargarse en su totalidad.

1.2 Corpus orales disponibles en su totalidad

En este apartado listamos los corpus orales cuyos textos se pueden obtener en su totalidad (mediante compra o descargándolos en internet), sin que sean consultables en internet. Este tipo de corpus permite un acceso directo a los textos, que pueden procesarse mediante cualquier equipo lógico propio para el análisis de corpus.

Corpus C-ORAL-ROM (parte portuguesa)

- <http://metashare.metanet4u.eu/repository/browse/c-oral-rom_exm/362a2020cf5711e1a404080027e73ea28eaaf998e9aa47739841451ea4e16f51/>; en venta en <http://catalog.elra.info/product_info.php?products_id=757>
- 300.000 palabras
- CLUL
- PE

El corpus C-ORAL-ROM (###1 Anthologies et corpus pan-romans) es un corpus comparable de lenguaje hablado para 4 idiomas romances (it., fr., ptg. y esp.), con las mismas dimensiones y constitución interna (Bacelar do Nascimento et al. 2005). Las grabaciones están transcritas en formato CHAT y la alineación entre sonido y transcripción se hizo con el programa WinPITCH. El corpus está etiquetado con etiquetas morfosintácticas POS. En la plataforma META-SHARE está disponible una nueva versión de la parte ptg., con revisión de la transcripción y alineación, realizadas con el programa EXMARaLDA, en formato XML.[5]

Corpus Português Fundamental

- <http://www.clul.ul.pt/en/resources/84-spoken-corpus-qportugues-fundamental-pfq-r>; nueva versión en venta en ELRA (gratuito para investigadores): <http://catalog.elra.info/product_info.php?products_id=1173>
- 700.000 palabras
- CLUL
- PE

[5] El programa EXMARaLDA permite la búsqueda de concordancias con audición del contexto seleccionado.

El Corpus *Português Fundamental* se compone de 1.800 grabaciones de conversaciones (500 horas, archivadas en el CLUL), realizadas en situaciones de comunicación oral espontánea, sobre gran variedad de temas cotidianos, con hablantes de edades, clases sociales y profesionales muy diversas (Bacelar do Nascimento/Garcia Marques/Segura da Cruz 1987; Bacelar do Nascimento/Rivenc/Segura da Cruz 1987). De estas conversaciones, 1.400 extractos fueron transcritos (sumando 700.000 palabras), que constituyen el llamado *Corpus de Frequência*. Este corpus fue uno de los primeros corpus de habla disponibles para el portugués, por lo que muchos estudios sobre el lenguaje hablado se basaron en él. En la nueva versión disponible en el catálogo ELRA, el corpus está anotado con POS, se presenta en formato XML, y texto y sonido pueden visualizarse alineados mediante el programa EXMARaLDA.

Corpus Português Falado: documentos autênticos

- <http://www.clul.ul.pt/pt/recursos/83-spoken-portuguese-geographical-and-social-varieties-r>; nueva versión en venta en el catálogo ELRA (gratuito para investigadores): <http://catalog.elra.info/product_info.php?products_id=1172>
- 92.000 palabras, de las cuales 30 textos para el PE
- CLUL
- PE y otras variantes nacionales

El corpus *Português Falado* forma un corpus oral de variantes nacionales del ptg. Contiene 86 grabaciones de conversaciones informales entre conocidos, amigos y familiares, así como intervenciones más formales como programas radiofónicos, de las décadas 1970, 1980 y 1990 para el PE (cf. Bacelar do Nascimento 2001). Los hablantes son de procedencia sociolingüística diversa y tienen el ptg. como lengua materna o segunda lengua. La caracterización de los hablantes aparece al inicio de cada transcripción (origen, sexo, edad, profesión, nivel de instrucción, información sobre las condiciones de la grabación como el local y la fecha). La primera versión se presenta en formato TXT para las transcripciones y las grabaciones en formato WAV y puede procesarse con editores de texto o audio comunes. También se puede descargar el programa *Lingua* que procesa ambos tipos de ficheros de forma alineada. La nueva versión del corpus está morfosintácticamente anotada (POS) y las transcripciones alineadas fueron realizadas con el programa EXMARaLDA, en formato XML.

Corpus HESITA

- <http://lsi.co.it.pt/spl/hesitation/downloads.html>
- 27 h de grabación
- Instituto de Telecomunicações
- PE

El corpus *HESITA* recopila grabaciones y transcripciones manuales de eventos de habla con vacilaciones en telediarios portugueses. El corpus está etiquetado para las vacilaciones lingüísticas de acuerdo con el sistema PLS (*Pattern Labeling System*) con algunas adaptaciones (Candeias et al. 2013).

Portuguese Batoreo Corpus

- <http://www.language-archives.org/item/oai:childes.talkbank.org:Romance-Portuguese-Batoreo>
- Dimensión en número de palabras desconocida
- Hanna Batoréo
- PE

El Portuguese Batoreo Corpus reúne textos orales de lenguaje infantil y adulto en los años 1992-1993: consiste de dos narraciones elicitadas a base de una serie de imágenes, cada una contada por 30 adultos y 30 niños (cf. Batoréo 2000). La edad de los niños es de 5, 7 y 10 años (10 participantes por edad), los participantes adultos tienen entre 18 y 45 años. Las transcripciones, con metadatos sobre los informantes y algunas marcas de oralidad (pausa, repeticiones, contracciones, etc.), están disponibles en formato XML y CHAT a través de la plataforma CHILDES (cf. MacWhinney 2012; Wilkens/Villavicencio 2012; *Guide to Childes Manual – Romance Corpora*[6]), pero sin acceso a las grabaciones.

Acquisition of European Portuguese Databank (AcEP) (Subset 1-3 y 4)

- Descarga online de las transcripciones en <http://childes.psy.cmu.edu/> bajo «Database» > Transcripts – Media – XML Browsable Database > ficheros «Freitas» y «CCF»
- Dimensión en número de palabras desconocida
- M. João Freitas, Susana Correia, Teresa Costa et al., CLUL
- PE

El AcEP es un corpus de datos longitudinales espontáneos de lenguaje infantil de 12 niños portugueses entre 0 y 4 años a base de sesiones de grabación mensuales o quincenales en los años 1990 y 2000 (Freitas et al. 2012). Las grabaciones audio (MP3 y WAV) de 5 niños están libremente accesibles en los ficheros en la plataforma CHILDES. El acceso a las grabaciones audio y vídeo se concede tras petición a la coordinadora de la ACeP Maria João Freitas. Las transcripciones ortográficas y fonológicas están libremente accesibles online y se están sometidas a los estándares de CHILDES en formato XML y CHAT (cf. referencias *supra – Portuguese Batoreo Corpus*).

El AcEP contiene otros corpus y bases de datos, de las que mencionamos uno que estará disponible en breve en la misma plataforma CHILDES, pero que puede solicitarse a la autora Laetitia Almeida. Se trata de un corpus de datos longitudinales de 4 niños bilingües portugués-francés (Subset 4 de la AcEP).

2 Selección de textos

Para ilustrar la variabilidad del PE contemporáneo, presentamos una serie de extractos textuales de índole diversa. En la sección 2.1, confrontamos dos tipos de textos muy distintos: un texto literario (2.1.1),

[6] http://childes.psy.cmu.edu/manuals/08romance.pdf.

ejemplo de un lenguaje meticulosamente trabajado, y unos extractos de comunicación digital (2.1.2), que presentan un lenguaje más espontáneo y con características muy propias. La sección 2.2 incluye textos orales: en 2.2.1 confrontamos un ejemplo de habla informal con un contexto de habla más formal; en 2.2.2 ilustramos rasgos lingüísticos caracterizadores para las dos grandes zonas dialectales que suelen distinguirse en el territorio continental portugués.

2.1 Textos escritos

2.1.1 Texto literário

Como ejemplo de texto literario, tomamos un extracto de *O Evangelho segundo Jesus Cristo* de José de Sousa Saramago (1922-2010) (Saramago 1991). Único autor portugués galardonado con el premio Nobel de Literatura (1998), se considera como uno de los mayores escritores del PE del siglo XX. Presenta un estilo experimental muy característico. Independientemente de elementos estilísticos individuales propios al autor y al texto, el discurso literario en general se caracteriza por ser un texto escrito cuidadosamente preparado, planificado y trabajado, lo que permite una organización original y creativa del contenido y de las ideas.

(a) Morfosintaxis
El estilo de Saramago se caracteriza por frases muy largas. El autor hace abundante uso de comas donde según la norma se esperarían puntos y seguidos o comillas para delimitar frases y turnos conversacionales (B2 *[…] disse, Muito desgraçados somos nós […]*). Estas frases largas se caracterizan por una sintaxis compleja, con gran cantidad de frases subord. Se nota específicamente la abundante inserción de subord. de part. (A1 *finalmente chegado*, A2 *apagadas as últimas cintilações*), ger. (A3 *cantando louvores*) y construcciones de inf. (flex.) (B2 *praticarmos a parte*…). También se observa el uso del pluscpf. sintético, hoy día caracterizador de un estilo arcaizante, formal y escrito (A4 *deixara, ouvira, fora*). Mencionamos, por último, la anteposición del adj. con respecto al subst., que produce un efecto poético de énfasis y de subjetividad y que se vincula con la función estética del texto literario (A1-2 *longa separação*, A5 *breve sonolência*, B4 *providencial palanque*, B6 *filosófica reflexão*).

(b) Léxico
Junto con el refinamiento estilístico en el plano sintáctico, el vocabulario es refinado y trabajado (por ej. A2 *cintilação*, A5 *sonolência*, B4 *providencial*, B5 *em ânsias*), con uso abundante de adj. calificativos (cf. ejemplos *supra*).

> A. Rejubilava em sua alma, e a si mesmo dizia que este era, finalmente chegado, o derradeiro dia da longa separação, que amanhã, logo à primeira hora, quando, apagadas as últimas cintilações dos astros, apenas brilhar céu a estrela Boieira, porá pés ao caminho, cantando louvores ao Senhor que nos guarda a casa e guia os passos
> Abriu de repente os olhos, sobressaltado, crendo que se deixara adormecer e não ouvira o sinal, mas fora apena

uma breve sonolência, os companheiros estavam aí todos, uns conversando, dormitando outros, e o manajeiro tranquilo, como se tivesse resolvido dar feriado aos seus operários e não pensasse arrepender-se da generosidad

B. O outro soldado, riscando o chão com o coto da lança, como o destino que parte e reparte, disse, Muito desgraçados somos nós, que não nos chega praticarmos a parte de mal que nos coube por natureza, e ainda temos de ser braço da maldade de outros e do seu poder. Estas palavras já não foram ouvidas por José, que se afastara do seu providencial palanque, primeiro de mansinho, pé ante pé, logo numa louca corrida, saltando as pedras como um cabrito, em ânsias, razão por que, faltando o seu testemunho, seja lícito duvidar da autenticidade da filosófica reflexão, quer quanto ao fundo quer quanto à forma, tendo em conta a mais do que óbvia contradição entre a notável propriedade dos conceitos e a ínfima condição social de quem os teria produzido.

2.1.2 Texto de comunicación mediada por ordenador

Presentamos aquí unos extractos de comunicación mediada por ordenador (en este caso, conversaciones de Facebook, sacados de un subcorpus del CRPC en construcción, con textos de blogs, Facebook, *tweets*, y otros tipos). Medio reciente y de alcance universal, la Red ha generado una amplia gama de nuevas formas de comunicación (foros, blogs, chat, e-mail, Twitter, etc.) en constante y rápida evolución. Por un lado, los diversos géneros en la Red se sitúan en un continuo con características tanto de lenguaje escrito como de lenguaje hablado. Por el otro lado, presentan rasgos propios que los distinguen de géneros textuales más establecidos y que a menudo resultan de las particularidades técnicas del medio (Crystal 2006). Así, pese a que la comunicación digital presenta turnos de conversación – a semejanza de la interacción verbal directa en el habla – difiere de esta por la ausencia de prosodia y de gestos paralingüísticos. Además, media cierto lapso de tiempo entre los turnos de conversación, que varía de menos de unos segundos a varios meses o incluso más.

(a) Características gráficas

El juego con los elementos gráficos compensa la falta de prosodia y de gestos con función expresiva. Así se notará el uso de mayúsculas para reflejar discursos en voz alta o para expresar indignación (A), tanto como el uso de puntos suspensivos (que pueden ser los tradicionales tres (D), pero a menudo son más de tres o solo dos (A)) para expresar pausas o continuación de un turno de conversación anterior (C). Muy llamativo, por supuesto, es el uso de emoticonos: en B surge uno que expresa vergüenza o escándalo, C contiene el «universal» emoticono riente, y el asterisco en E representa un beso de despedida. Además, la índole particular del medio entre conversación (rápida) y escrita (lenta) inspira un uso económico de la ortografía y cierta falta de cuidado ortográfico: los extractos ilustran la omisión de acentos ortográficos (A *pontape*, B *ja*, *nao*, E: *so*), de espacios (A *?LOL ..*, E: *gostam!!Beijo*), de signos de puntuación (A *kerias o ke amarelo?* para ‘querias o quê**,** amarelo?’) y de mayúsculas donde según las reglas del discurso escrito se esperarían

(B *ja*, D *com*, E *facebook*), así como un uso creativo de la ortografía para representar más económicamente ciertos sonidos (A qu > k en *kerias o ke* = 'querias o quê', B *1x*).

(b) Morfosintaxis
La conversación en la Red se caracteriza por una complejidad morfosintáctica flexible, que puede ser más espontánea y parecida al habla, o más conforme al lenguaje escrito. Así, los fragmentos A hasta E representan distintos grados de complejidad: A, B y C contienen frases más breves y giros típicamente hablados (A la pregunta entonativa *querias o quê, amarelo?* por *O que é que querias, amarelo?*), mientras que E presenta un texto más elaborado con frases relativamente más largas, conformes a los estándares escritos y sintácticamente más complejas (por ej. con subord.: *só me resta desejar um maravilhoso natal…*, … *rodeados de quem mais gostam*).

(c) Léxico
La conversación internáutica se caracteriza por la alternancia de códigos, en particular en cuanto al uso de expresiones de jerga ingl., muchas veces en forma de abreviaturas (A *LOL* = 'laughing out loud'). Además, los enunciados más próximos al lenguaje hablado llevan al uso de vocabulario informal (A *gajo* por *homem*).

A. A 4 MIN DO FIM ?LOL ..CADOZO MANDA UM PONTAPE NO GAJO E KERIAS O KE AMARELO?

B. -.- ja mandaste 1x a piada. nao te chegou ?

C. ... é altura disto e de muitas mais coisas!:)

D. com tanta importância que lhe dão, ela deve ter com certeza um bom «padrinho»... mas isso sou eu que acho!

E. Nem sei que dizer para exprimir todo o carinho destes dois dias. Obrigado a todas as centenas e centenas de pessoas que me deram os parabéns pelo facebook e das muitas e muitas mensagens privadas! Obrigado! So me resta desejar um maravilhoso natal rodeados de quem mais gostam!!Beijo e abraço*

2.2 Textos de oralidad

2.2.1 El habla formal e informal

Se presentan aquí dos textos característicos de dos tipos de producción oral, uno informal y otro formal, sacados del C-ORAL-ROM ptg. (véase *supra*). El texto A ejemplifica la producción oral informal: es una conversación entre dos jóvenes adultos y amigos, en la que X le relata a Y su viaje a Sarajevo. El texto B forma un extracto de una clase de enseñanza superior: ilustra el discurso público de índole formal.

(a) Prosodia
La oralidad presenta rasgos propios de una producción lingüística cuyo procesamiento es simultáneo con la enunciación. En consecuencia, contrasta con la producción escrita, que permite la revisión del texto y la consecuente eliminación de reformulaciones. Este procesamiento en curso lleva a la ocurrencia de algunos

rasgos llamativos del discurso oral informal no preparado: pausas rellenadas (A4 *ah*), hesitaciones (A2 *dá / dá-te*), reformulaciones (A2-3 *vais p* se reformula por *em Lisboa*) y frecuentes muletillas (A *pá*, *e não sei quê, estás a ver*). Obsérvese, así, la mayor frecuencia de estos fenómenos en el texto A, en comparación con el texto B, de registro formal y más preparado.

(b) Morfosintaxis

Los interlocutores del texto A, por su relación de amistad, por el carácter informal de la interacción y por su edad, usan la forma de tratamiento de pers.2 (A4 *percebes*), que no suele utilizarse en contextos formales.

Por lo que respecta a la estructuración del discurso, el texto A progresa sobre todo mediante el uso de conj. coord. y marcadores discursivos (*de repente*, *depois*, *então*) y mediante la yuxtaposición de secuencias: A2-3 *tu andas aqui a atrofiar // vais p/ em Lisboa andas lá todo / deprimido*; A3-4 *estas pessoas a família morreu / ah / percebes // a casa ruiu*; A9 *é a cidade toda / é só campas*. El texto B presenta una cohesión textual en el encadenamiento de la información que sugiere cierto grado de preparación de los contenidos de la clase.

(c) Léxico y discurso

En el texto A, la edad relativamente joven de los interlocutores se refleja, por ej., en el uso de la palabra *estilo* como forma de señalar una ejemplificación (A8-9 *estilo / nos quintais / nos jardins*), en el uso de la expresión *do género*, que, entre largas pausa, tiene la función de introducir la continuación del discurso, en las elecciones léxicas (A3 *atrofiar*) o en el uso de extranjerismos (A1 *power*). En cambio, el texto B no presenta léxico de registro familiar. Además, la situación discursiva específica de la enseñanza superior hace que se emplee gran número de términos especializados: B6-7 *métodos de Lagrange-Newton*, B7-8 *métodos de programação quadrática sequencial*.

A.[7] *X: é esse power / pá / é muito estranho // é muito estranho // mas ao mesmo tempo dá / dá-te um baque // mesmo ... naquela / o quê ? bem / tu andas aqui a atrofiar // vais p / em Lisboa andas lá todo / deprimido / e não sei quê // de repente estas pessoas a família morreu / ah / percebes // a casa ruiu // e / estão aqui / cheios de energia // não têm emprego / e não sei quê // pá / depois é / é / outra coisa q / que é impressionante na cidade é os cemitérios por todo o lado // porque aquilo durante a guerra / os cemitérios da cidade / eram no ja / nos limites onde estavam os / os inimigos // então eles tinham de enterrar os mortos à noite / estilo / nos quintais / nos jardins // então tu / é a cidade toda / é só campas // do género // as cenas mais impressionantes foi assim b / bairros de casas d / de piso térreo / estás a ver // com /
*Y: hum hum //
*X: / com quintais // porque em vez de quintal / é um cemitério // mesmo com aquelas / barrazinhas / de metal / estás a ver // mas que lá dentro está cemitério // pá / típico quintal / estás a ver // <mas com / lápides> //
*Y: [<] <mas / mas junto da casa> ?

[7] Optamos por mantener las marcas de pausa breve (/) y larga (//) para recuperar información prosódica.

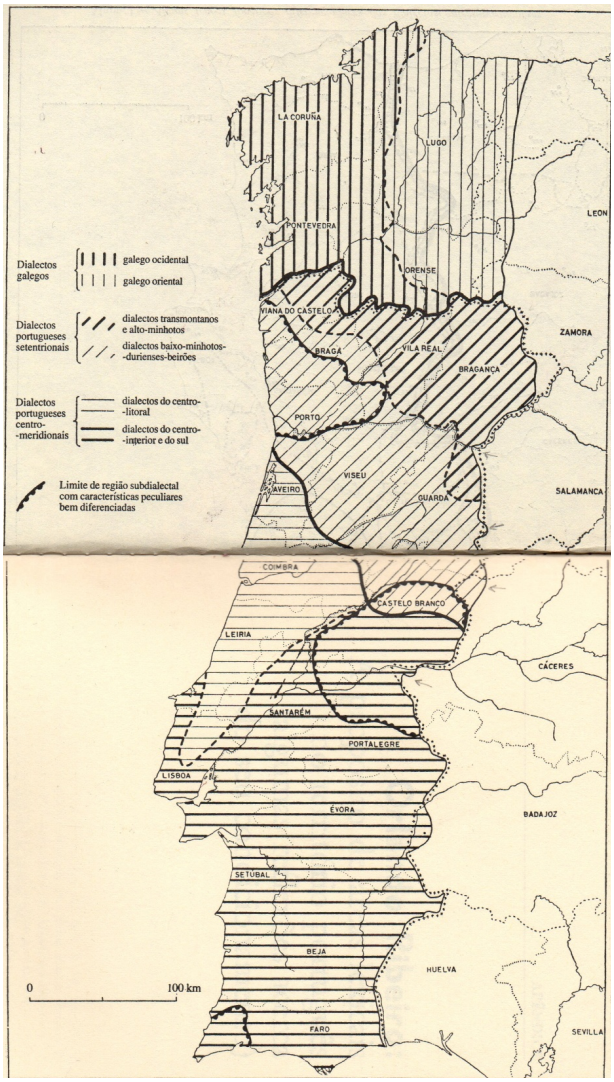
*X: sim // mesmo ao lado de <os cemitérios>
(Corpus C-Oral-Rom: informal, família/privado, monólogo)

B. *FER: portanto o nosso sumário hoje é um tanto ou quanto extenso // corresponde digamos / ao conteúdo de um capítulo novo // este capítulo / cujo sumário / está aqui no quadro / e que / aguardarei / e darei o tempo para que possam / ah / passar para os vossos cadernos // este último sumário / correspondente ao / nosso último capítulo / de acordo com o programa que foi distribuído / no início do semestre / fala / de / os chamados métodos de Lagrange-Newton // ou / como são também mais / frequentemente conhecidos / por métodos / de programação quadrática sequencial // ah / ou / pela sua sigla / SQP // isto é / sequencial quadratic programming methods // portanto métodos de programação quadrática sequencial // ah / este po / a questão que se põe é / estes métodos são novos ? não // estes métodos na sua ideia / são até relativamente antigos // e já / perceberão / a seguir / porquê //
(Corpus C-Oral-Rom: formal, aula no ensino superior)

2.2.2 Variación dialectal

A pesar de la relativa homogeneidad del idioma ptg. en el territorio portugués, se pueden distinguir los dial. septentrionales, centro-meridionales y los dial. insulares (Madeira y Azores). Esta tripartición se basa en criterios fonéticos y fonológicos, aunque también hay rasgos diferenciadores en el plano morfológico, sintáctico y léxico (Segura 2013, Vasconcellos 1970). El Mapa 1, que reproduce un mapa de Lindley Cintra (Cintra 1983b, 162s.), marca la división territorial entre los dial. septentrionales y centromeridionales (y además los dial. gallegos). Las características de los dial. centro-meridionales coinciden generalmente con aquellas de la variedad estándar del PE (a la excepción de la monoptongación del diptongo [ej], como veremos con respecto a las características fonéticas y fonológicas). A fin de ilustrar estas diferencias dialectales, presentamos dos transcripciones de grabaciones del corpus ALEPG (cf. *supra*), seleccionadas por Luísa Segura, miembro diste proyecto. El texto A, producido por un hablante de Cabril, Viana do Castelo, es ilustrativo de los dialectos septentrionales; el texto B, de un hablante de Corte Cobres, Beja, ilustra los dialectos centro-meridionales.[8]

[8] Agradecemos a Luísa Segura su valiosa ayuda con la selección de las dos transcripciones del ALEPG y con los comentarios de sus características dial.



Mapa 1: clasificación de los dialectos gallego-portugueses (Cintra 1983b, 162s.)

(a) Fonética y fonología

Los dial. septentrionales muestran varios rasgos distintivos, de los que ilustramos aquí dos: la neutralización fonológica entre /v/ y /b/, a favor del /b/, pronunciada generalmente como bilabial fricativa [β] (A1-2 *vós*, *-vos*, *vosso*, *vos*, *vou*) y la conservación del diptongo [ow] (A2 *mandou*, *vou*). En los dial. centro-meridionales, señalamos la monoptongación del diptongo [ej] hacia [e] (B3 *maneiras*, B4 *alqueires*).

(b) Morfosintaxis y sintaxis

En los dial. del norte se mantiene el uso del pronombre de pers.5 *vós* y de las correspondientes formas verbales (A1 *vós*, *inde-vos*[9]). En los dial. centro-meridionales y en la variedad de prestigio, la pers.5 se substituye generalmente por la forma *vocês*, que representa gramaticalmente una pers.6, con la correspondiente flexión verbal. Además, en las zonas septentrionales, las formas de la conj. II (con desinencia en *-er*) se sustituyen a veces por *-ende*, *-endes* bajo la influencia del V *ter* (*tende*, *tendes*). Así, en el texto A, la forma *dizende* sustituye la pers.5 del imp. *dizei* (A1).

A menudo, los dial. muestran tendencia a regularizar la conj. verbal. Los dial. centro-meridionales, por ejemplo, presentan casos en que la forma de la pers.1 del pret. ind. de los V de la conj I (en *-ar*) se termina en *-i* en vez de en *-ei* (B1 *lidei* se pronuncia como *lidi*).

Por último, los dial. septentrionales se caracterizan por la frecuente inserción de constituyentes entre el pron. pers. átono antepuesto y el V. El texto A ilustra un caso de inserción de un adv. (A2 *vos cá mandou*).

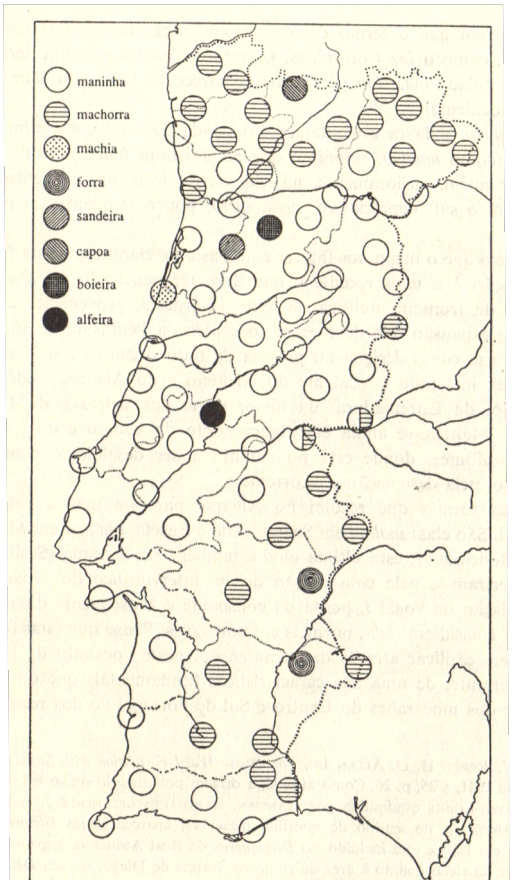
(c) Léxico

En 1962 Lindley Cintra propuso una división del territorio portugués en distintas áreas léxicas (Cintra 1983b). Una primera división opone los dial. septentrionales a los meridionales: aquellos manifiestan a veces vocablos de origen lat. o germánico, donde estos emplean palabras de origen ár. para el mismo concepto. Es el caso, por ejemplo, de *soro* (lat.) y *almece* (ár.), respectivamente, que designan ambos el líquido que se separa de la leche al cuajar (cf. Mapa 2: Cintra 1983a, 69). Otra división léxica contrapone un área conservadora en el noroeste y el oeste del país (hasta la zona norte de la provincia Estremadura) al área innovadora en las zonas al este y al sur del país: por ej., *moger/mugir* (noroeste) vs. *ordenhar* (este y sur).

> A. X – E o gajo disse assim: «E vós inde-vos embora (…) E dizende lá ao vosso amigo, ao que vos cá mandou que venha cá ele; que eu amanhã que já vou a Lisboa resolver o problema dos espanhóis, porque os nossos portugueses quando foi da Grande Guerra também fugiram por lá e ninguém os prendeu nem ninguém lhes fez mal e muitos ainda lá vivem hoje» (ALEPG, Cabril, Vila Real)

[9] En este caso, la forma verbal presenta además nasalización de la vocal (forma normativa: *ide*).

B. X – Bom, lidei ali com as ovelhas até, até que fui para a tropa. (…) Quando voltei da tropa. Os patrões vieram-me ver, disseram-me logo «Tens aqui a casa, tal e qual, em querendes vir…». (…) De maneiras que davam-me essas 40 cabecinhas, com as borregas eram 50 e davam-me 20 alqueires de trigo e davam-me 100 mel-
5 réis por ano. (ALEPG, Corte Cobres, Beja)



Mapa 2: Áreas de expansión de los sustantivos *soro* y *almece* en el territorio portugués (Cintra 1983a, 69)

3 Bibliografía

Bacelar do Nascimento, M. Fernanda (2000), *Corpus de Référence du Portugais Contemporain*, in: Mireille Bilger (ed.), *Corpus, Méthodologie et Applications Linguistiques*, Paris, Champion/Presses Universitaires de Perpignan, 25-30.

Bacelar do Nascimento, M. Fernanda (coord.) (2001), *Português Falado, Documentos Autênticos, Gravações audio com transcrições alinhadas*, Lisboa, Centro de Linguística da Universidade de Lisboa e Instituto Camões [cederrón].

Bacelar do Nascimento, M. Fernanda/Garcia Marques, M. Lúcia/Segura da Cruz M. Luísa (1987), *Português Fundamental, Métodos e Documentos*, tomo 1: *Inquérito de Frequência*, Lisboa, INIC, CLUL.

Bacelar do Nascimento, M. Fernanda/Rivenc, Paul/Segura da Cruz, M. Luísa (1987), *Português Fundamental, Métodos e Documentos*, tomo 2: *Inquérito de Disponibilidade*, Lisboa, INIC, CLUL.

Bacelar do Nascimento, M. Fernanda, et al. (1998), *LE-PAROLE – Do corpus à modelização da informação lexical num sistema-multifunção*, in: *Actas do XIII Encontro da Associação Portuguesa de Linguística*, Lisboa, APL, 115-134.

Bacelar do Nascimento, M. Fernanda, et al. (2005), *The Portuguese Corpus*, in: Emanuela Cresti/Massimo Moneglia (edd.), *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, Amsterdam/Philadelphia, Benjamins, 163-207 [con cederrón].

Barreto, Florbela, et al. (2006), *Open Resources and Tools for the Shallow Processing of Portuguese*, in: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, Genova, Italia, 1438-1443.

Batoréo, Hanna Jakubowicz (2000), *Expressão do Espaço no Português Europeu. Contributo psicolinguístico para o Estudo da Linguagem e Cognição*, tesis doctoral, Lisboa, Fundação Calouste Gulbenkian/Fundação para a Ciência e a Tecnologia.

Candeias, Sara, et al. (2013), *HESITA(tions) in Portuguese: a database*, in: *Proceedings of the 6th workshop on Disfluency in Spontaneous Speech*, Stockholm, KTH Royal Institute of Technology, 13-16.

Carrilho, Ernestina (2010), *Tools for dialect syntax: the case of CORDIAL-SIN (an annotated corpus of Portuguese dialects)*, in: Gotzon Aurrekoetxea/Jose Luis Ormaetxea (edd.), *Tools for Linguistic Variation*, Bilbao, Universidad del País Vasco, 57-70.

Cintra, Luís F. Lindley [1962] (1983a), *Áreas lexicais no território português*, in: *Estudos de Dialectologia Portuguesa*, Lisboa, Sá da Costa, 55-94.

Cintra, Luís F. Lindley [1971] (1983b), *Nova proposta de classificação dos dialectos galego-portugueses*, in: *Estudos de Dialectologia Portuguesa*, Lisboa, Sá da Costa, 117-163.

Crystal, David (2006), *Language and the internet*, Cambridge, Cambridge University Press.

Davies, Mark (2014), *Creating and using the Corpus do Português and the Frequency Dictionary of Portuguese*, in: Tony Berber Sardinha/Telma Ferreira (edd.), *Working with Portuguese Corpora*, London, Continuum, 89-110.

Freitas, Cláudia/Afonso, Susana (2008), *Bíblia Florestal: Um manual lingüístico da Floresta Sintá(c)tica*, <http://linguateca.dei.uc.pt/Floresta/BibliaFlorestal/> (27.03.2014).

Freitas, Maria João, et al. (2012), *Child-Adult Interaction: A Database on European Portuguese*, Lisboa, CLUL, Anagrama.

Généreux, Michel/Hendrickx, Iris/Mendes, Amália (2012), *Introducing the Reference Corpus of Contemporary Portuguese On-Line*, in: *Proceedings of the Eighth International Conference on Language Resources and Evaluation – LREC 2012,* Istanbul, ELRA, 2237-2244.

MacWhinney, Brian (2012), *The CHILDES Project. Tools for Analyzing Talk - Electronic Edition*, part 1: *The CHAT Transcription Format*, <http://childes.psy.cmu.edu/manuals/CHAT.pdf> (27.03.2014).

Rocha, Paulo/Santos, Diana (2000), *CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa*, in: Maria das Graças Volpe Nunes (ed.), *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*, Atibaia, São Paulo, Brasil, 131-140.

Rocha, Paulo/Santos, Diana (2007), *Disponibilizando a Colecção Dourada do HAREM através do projecto AC/DC*, in: Diana Santos/Nuno Cardoso (edd.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, Linguateca, 307-326.

Saramago, José (1991), *O Evangelho segundo Jesus Cristo*, Lisboa, Caminho.

Segura, Luísa (2013), *Variedades dialectais do português europeu*, in: Eduardo B. Paiva Raposo et al. (edd.), *Gramática do Português*, vol. I, Lisboa, Fundação Calouste Gulbenkian, 85-142.

Silva, Augusto Soares da (2008), *O corpus CONDIV e o estudo da convergência e divergência entre variedades do português*, in: Luís Costa/Diana Santos/Nuno Cardoso (edd.), *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, Linguateca, 25-28.

Vasconcellos, J. Leite de (1970), *Esquisse d'une dialectologie*, Lisboa, Centro de Estudos Filológicos.

Wilkens, Rodrigo/Villavicencio, Aline (2012), *Uma ferramenta para a pesquisa em corpora de aquisição de linguagem*, in: Póster del *XI Encontro de Linguística de Corpus*, São Carlos, Brasil, <http://www.nilc.icmc.usp.br/elc-ebralc2012/poster/104055.pdf> (27.03.2014).

CLARA VANDERSCHUEREN y AMÁLIA MENDES